AISSMS
INSTITUTE OF INFORMATION TECHNOLOGY
(IOIT)
ADDING VALUE TO ENGINEERING
An Autonomous Institute Affiliated to Savitribai Phule Pune University
Approved by AICTE, New Delhi and Recognised by Govt. of Maharashtra
Accredited by NAAC with "A+" Grade | NBA - 5 UG Programmes

MINI PROJECT

Disney Movie Production Prediction

Submitted by

28 – SAGAR JADHAV

29 – PRAJWAL JOSHI

34 – RUTWIK KHANDGAWHRE

Aim:

Should Disney make more action and adventure movies?

1. The dataset

Walt Disney Studios is the foundation on which The Walt Disney Company was built. The Studios has produced more than 600 films since their debut film, Snow White and the Seven Dwarfs in 1937. While many of its films were big hits, some of them were not. In this notebook, we will explore a dataset of Disney movies and analyze what contributes to the success of Disney movies.

First, we will take a look at the Disney data compiled by Kelly Garrett. The data contains 579 Disney movies with six features: movie title, release date, genre, MPAA rating, total gross, and inflation-

adjusted gross.

Let's load the file and see what the data looks like.

```
# Import pandas library
# ... YOUR CODE FOR TASK 1 ...
import pandas as pd
# Read the file into gross
gross = pd.read_csv("datasets/disney_movies_total_gross.csv")
gross['release_date'] = gross['release_date'].apply(pd.to_datetime)

# Print out gross
gross.head()
```

	movie_title	release_date	genre	mpaa_rating	total_gross	inflation_
0	Snow White and the Seven Dwarfs	1937-12-21	Musical	G	184925485	
1	Pinocchio	1940-02-09	Adventure	G	84300000	
2	Fantasia	1940-11-13	Musical	G	83320000	
3	Song of the South	1946-11-12	Adventure	G	65000000	
4	Cinderella	1950-02-15	Drama	G	85000000	

2. Top ten movies at the box office

Let's started by exploring the data. We will check which are the 10 Disney movies that have earned the most at the box office. We can do this by sorting movies by their inflation-adjusted gross (we will call it adjusted gross from this point onward).

```
# Sort data by the adjusted gross in descending order
# ... YOUR CODE FOR TASK 2 ...
gross=gross.sort_values(by="inflation_adjusted_gross",ascending=False)
# Display the top 10 movies
# ... YOUR CODE FOR TASK 2 ..
gross.head(10)
```

	movie_title	release_date	genre	mpaa_rating	total_gross	inflation_
0	Snow White and the Seven Dwarfs	1937-12-21	Musical	G	184925485	
1	Pinocchio	1940-02-09	Adventure	G	84300000	
2	Fantasia	1940-11-13	Musical	G	83320000	
8	101 Dalmatians	1961-01-25	Comedy	G	153000000	

3. Movie genre trend

From the top 10 movies above, it seems that some genres are more popular than others. So, we will check which genres are growing stronger in popularity. To do this, we will group movies by genre and then by year to see the adjusted gross of each genre in each year.

```
# Extract year from release_date and store it in a new column
gross['release_year'] = pd.DatetimeIndex(gross['release_date']).year

# Compute mean of adjusted gross per genre and per year
group = gross.groupby(['genre','release_year']).mean()

# Convert the GroupBy object to a DataFrame
genre_yearly = group.reset_index()

# Inspect genre_yearly
genre_yearly.head(10)
print(genre_yearly.info())
```

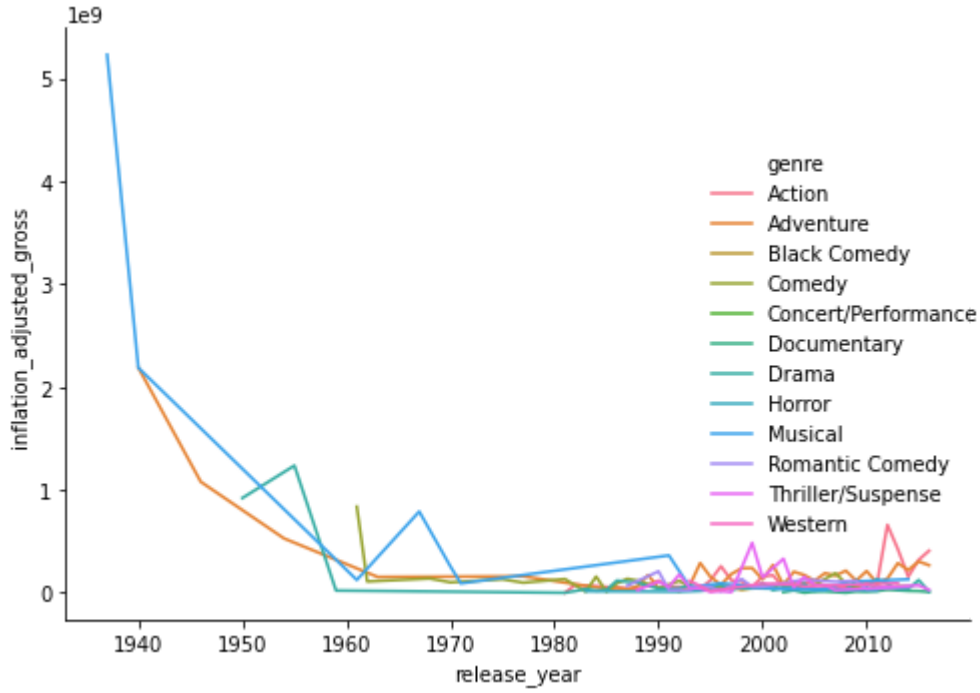
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 218 entries, 0 to 217
Data columns (total 4 columns):
genre                       218 non-null object
release_year                218 non-null int64
total_gross                 218 non-null float64
inflation_adjusted_gross    218 non-null float64
dtypes: float64(2), int64(1), object(1)
memory usage: 6.9+ KB
None
```

4. Visualize the genre popularity trend

We will make a plot out of these means of groups to better see how box office revenues have changed over time.

```
# Import seaborn library
# ... YOUR CODE FOR TASK 4 ...
import seaborn as sns
# Plot the data
sns.relplot(data=genre_yearly,x='release_year',y='inflation_adjusted_gross',kind='line',hue='
# ... YOUR CODE FOR TASK 4 ...
```

```
<seaborn.axisgrid.FacetGrid at 0x7f63955149b0>
```



5. Data transformation

The line plot supports our belief that some genres are growing faster in popularity than others. For Disney movies, Action and Adventure genres are growing the fastest. Next, we will build a linear regression model to understand the relationship between genre and box office gross.

Since linear regression requires numerical variables and the genre variable is a categorical variable, we'll use a technique called one-hot encoding to convert the categorical variables to numerical. This technique transforms each category value into a new column and assigns a 1 or 0 to the column.

For this dataset, there will be 11 dummy variables, one for each genre except the action genre which we will use as a baseline. For example, if a movie is an adventure movie, like The Lion King, the adventure variable will be 1 and other dummy variables will be 0. Since the action genre is our baseline, if a movie is an action movie, such as The Avengers, all dummy variables will be 0.

```
# Convert genre variable to dummy variables
genre_dummies = pd.get_dummies(gross['genre'],drop_first=True)

# Inspect genre_dummies
```

```
genre_dummies.head()
genre_dummies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 579 entries, 0 to 29
Data columns (total 11 columns):
Adventure              579 non-null uint8
Black Comedy           579 non-null uint8
Comedy                 579 non-null uint8
Concert/Performance    579 non-null uint8
Documentary            579 non-null uint8
Drama                  579 non-null uint8
Horror                 579 non-null uint8
Musical                579 non-null uint8
Romantic Comedy        579 non-null uint8
Thriller/Suspense      579 non-null uint8
Western                579 non-null uint8
dtypes: uint8(11)
memory usage: 10.7 KB
```

6. The genre effect

Now that we have dummy variables, we can build a linear regression model to predict the adjusted gross using these dummy variables.

From the regression model, we can check the effect of each genre by looking at its coefficient given in units of box office gross dollars. We will focus on the impact of action and adventure genres here. (Note that the intercept and the first coefficient values represent the effect of action and adventure genres respectively). We expect that movies like the Lion King or Star Wars would perform better for box office.

```
# Import LinearRegression
# ... YOUR CODE FOR TASK 6 ...
from sklearn.linear_model import LinearRegression
# Build a linear regression model
regr = LinearRegression()

# Fit regr to the dataset
# ... YOUR CODE FOR TASK 6 ...
regr.fit(genre_dummies,gross['inflation_adjusted_gross'])
# Get estimated intercept and coefficient values
action =  regr.intercept_
adventure = regr.coef_[[0]][0]

# Inspect the estimated intercept and coefficient values
print((action, adventure))
```

```
(102921757.36842026, 87475654.70909917)
```

7. Confidence intervals for regression parameters (i)

Next, we will compute 95% confidence intervals for the intercept and coefficients. The 95% confidence intervals for the intercept ***a*** and coefficient b_i means that the intervals have a probability of 95% to contain the true value ***a*** and coefficient b_i respectively. If there is a significant relationship between a given genre and the adjusted gross, the confidence interval of its coefficient should exclude 0.

We will calculate the confidence intervals using the pairs bootstrap method.

```
# Import a module
import numpy as np

# Create an array of indices to sample from
inds = np.arange(0,len(gross['genre']))

# Initialize 500 replicate arrays
size = 500
bs_action_reps =  np.empty(size)
bs_adventure_reps =  np.empty(size)
```

8. Confidence intervals for regression parameters (ii)

After the initialization, we will perform pair bootstrap estimates for the regression parameters. Note that we will draw a sample from a set of (genre, adjusted gross) data where the genre is the original genre variable. We will perform one-hot encoding after that.

```
# Generate replicates
print(gross.info())

for i in range(size):

    # Resample the indices
    bs_inds = np.random.choice(inds,size=len(inds))

    # Get the sampled genre and sampled adjusted gross
    bs_genre = gross['genre'][bs_inds]
    bs_gross = gross['inflation_adjusted_gross'][bs_inds]

    # Convert sampled genre to dummy variables
    bs_dummies = pd.get_dummies(bs_genre,drop_first=True)

    # Build and fit a regression model
    regr = LinearRegression().fit(bs_dummies, bs_gross)
```

```
    # Compute replicates of estimated intercept and coefficient
    bs_action_reps[i] = regr.intercept_
    bs_adventure_reps[i] = regr.coef_[[0]][0]
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 579 entries, 0 to 29
Data columns (total 7 columns):
movie_title                 579 non-null object
release_date                579 non-null datetime64[ns]
genre                       562 non-null object
mpaa_rating                 523 non-null object
total_gross                 579 non-null int64
inflation_adjusted_gross    579 non-null int64
release_year                579 non-null int64
dtypes: datetime64[ns](1), int64(3), object(3)
memory usage: 36.2+ KB
None
```

9. Confidence intervals for regression parameters (iii)

Finally, we compute 95% confidence intervals for the intercept and coefficient and examine if they exclude 0. If one of them (or both) does, then it is unlikely that the value is 0 and we can conclude that there is a significant relationship between that genre and the adjusted gross.

```
# Compute 95% confidence intervals for intercept and coefficient values
confidence_interval_action = np.percentile(bs_action_reps,[2.5,97.5])
confidence_interval_adventure = np.percentile(bs_adventure_reps,[2.5,97.5])

# Inspect the confidence intervals

print(confidence_interval_action)
print(confidence_interval_adventure)
```

```
[7.22857233e+07 1.41582371e+08]
[3.47867244e+07 1.44172135e+08]
```

10. Should Disney make more action and adventure movies?

The confidence intervals from the bootstrap method for the intercept and coefficient do not contain the value zero, as we have already seen that lower and upper bounds of both confidence intervals are positive. These tell us that it is likely that the adjusted gross is significantly correlated with the action and adventure genres.

From the results of the bootstrap analysis and the trend plot we have done earlier, we could say that Disney movies with plots that fit into the action and adventure genre, according to our data, tend to

do better in terms of adjusted gross than other genres. So we could expect more Marvel, Star Wars, and live-action movies in the upcoming years!

```
# should Disney studios make more action and adventure movies?
more_action_adventure_movies = True
```